DOCUMENT RESUME

ED 187 768 TM 800 345

AUTHOR Haney, Walt
TITLE Testing the Tests. National Consortium on Testing Staff Circular No. 1.
INSTITUTION National Consortium on Testing, Cambridge, Mass.
PUB DATE Feb 78
NOTE 14p.
AVAILABLE FROM Huron Institute, National Consortium on Testing Project, 123 Mt. Auburn St., Cambridge, MA 02138 ($2.00)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Educational Testing; Elementary Education; Evaluation Criteria; *Standardized Tests; *Standards; Testing Problems; *Test Reviews
IDENTIFIERS Center for the Study of Evaluation CA; Mean Test Evaluation System; Mental Measurements Yearbook; *Test Bibliographies

ABSTRACT

This paper discusses the quality of tests with respect to their different uses. Standardized tests are defined as standardized procedures for eliciting responses from individuals, and these form the basis for some sort of appraisal. They can include intelligence tests, achievement tests, school readiness tests, and self concept measures. Oscar Buros' Series of Mental Measurements Yearbooks include critical reviews of tests, but these test reviews represent many different viewpoints, making it difficult to form general conclusions. The University of California, Los Angeles' Center for the Study of Evaluation (CSE) has published a series of test evaluations. Their Elementary School Test Evaluations are described in detail, including the MEAN evaluation system's criteria for judging tests: measurement validity, examinee appropriateness, administrative usability, and normal technical excellence. A brief review of the CSE evaluations suggests that very few tests are judged to be of high quality on all four criteria. A two-page list of information sources is appended. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED187768

The Huron Institute
NATIONAL CONSORTIUM ON TESTING PROJECT
123 MOUNT AUBURN STREET
CAMBRIDGE, MASSACHUSETTS 02138
617-661-9285

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

NATIONAL CONSORTIUM ON TESTING

STAFF CIRCULAR NO. 1

TESTING THE TESTS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. Haney
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

February 1978

Walt Haney

The Huron Institute
National Consortium on Testing Project
David K. Cohen, Executive Director
123 Mt. Auburn Street
Cambridge, MA 02138
(617) 661-9285

This is the first of a series of staff circulars which will be distributed as part of the Huron Institute's staff work for the National Consortium on Testing.

This first circular, "Testing the Tests," discusses efforts to assess the quality of standardized tests. This first circular is part of Huron's work on "Truth in Testing." The circular does not represent a comprehensive treatment of this area; rather it is an initial attempt to treat several quite limited questions regarding truth in testing in a condensed format. The circular is intended to spur discussion on this topic, and to raise important issues worthy of additional attention, not to resolve them. It is produced primarily as a resource document for member organizations of the NCT.

Although the circular has been prepared as part of staff work for the Consortium, needless to say, it does not necessarily reflect the views of individual Consortium members nor of the Consortium as a whole.

TM 800345

N A T I O N A L C O N S O R T I U M O N T E S T I N G

S T A F F C I R C U L A R N O . 1*

February 1978

TESTING THE TESTS

In exploring the quality and usefulness of tests, some natural first questions are:

- What are these things called "tests"?
- What are these tests good for and what are they not good for?
- How good are they for the functions they serve?
- How might these functions be better served?

These questions are simple, but answers to them are not. In this paper we review some evidence on the first three of these questions. Specifically, we briefly summarize evidence from efforts to evaluate tests themselves. We will deal again with these questions in later work.

The sort of tests that concern us here are standardized educational tests--specifically tests that might be defined as standardized procedures for eliciting responses from individuals, and that form the basis for some sort of appraisal. The things to be appraised vary: there are standardized tests of intelligence, of mathematics and reading achievement, of readiness for school, and even standardized tests of self-esteem. There are literally thousands of standardized tests of many different sorts. A recent review, for example, found more than 1,570 different test instruments in 3500 project proposals funded by 23 different federal agencies in fiscal year 1975.**

Oscar Buros--Granddaddy of Test Bibliographers

There are several sorts of reviews on standardized tests available. Probably the best known are a series of books by Oscar Buros, who is doubtless the granddaddy of all test bibliographers. For decades, Buros' Mental Measurements Yearbooks have been the most widely cited source of

*This circular has been prepared by staff of the National Consortium on Testing. The circular does not, however, necessarily reflect the views of the Consortium as a whole nor the views of individual Consortium members.

**Heyneman, Stephen P., Mintz, Pamela Cope. The Frequency and Quality of Measures Utilized in Federally-Sponsored Research on Children and Adolescents. Social Research Group of the George Washington University, Washington, D.C.: Spring 1976. The majority of tests identified in this review were unpublished tests.

information both on published tests and on critical reviews of them.* Indeed one of the prime objectives of the Mental Measurements Yearbooks (MMY) since the series was begun in 1938 has been to "present frankly critical test reviews written by testing and subject specialists representing various viewpoints." One problem in using the MMY's, however, is that the reviews of tests in them do indeed represent various viewpoints. So it is impossible to summarize these reviews in any straightforward manner to derive a general answer to the question of how good are standardized tests. Buros' own summary is given in his introduction to the Seventh MMY. He comments, "At least half of the tests currently on the market should never have been published. Exaggerated, false, or unsubstantiated claims are the rule rather than the exception (pp. xxvii-xxviii).

CSE Test Evaluations

Another source of information on published standardized tests offers a more systematic overview of their quality. This is a series of test evaluations published by the UCLA Center for the Study of Evaluation (CSE). The UCLA efforts to test the tests began in the late 1960 Since then the CSE group has published five different books of test evaluations.** The most recent volume is the CSE Elementary School Test Evaluations published in 1976. Since this book is the most recent of the CSE test evaluations, and presumably includes the most up-to-date tests, we focus attention here on this volume. For its review of elementary tests, the CSE acquired copies of some 800 published standardized tests, including over 3900 subtests which yielded separately interpretable test scores, according to test publishers' claims.

MEAN Test Evaluation System

The CSE evaluators tested the tests as follows. First, they categorized each subtest according to one of fourteen goal areas and also according to the grade level for which the test was designed. Then they evaluated each subtest according to what they called the MEAN test evaluation system. The acronym stands for: measurement validity; examinee appropriateness; administrative usability; and normed technical excellence.

Measurement validity was used by CSE to refer to publishers' evidence regarding the validity of what their tests measured. Included in this category was evidence regarding content, construct, concurrent and predictive validity of the tests. Examinee appropriateness was used to encompass information regarding the appropriateness of the tests with regard to students for whom they were intended. Ratings in this category were based on publishers' justifications of tests' appropriateness, the comprehension level of the tests, instructions for students taking the tests, test format, time and pacing of the tests and the manner in which answers were recorded. Administrative

*A brief bibliography of Buros' Mental Measurements Yearbooks and other volumes is appended to this circular.

**References to the CSE test evaluation series are appended to this circular.

usability was used by CSE to denote issues of how tests were administered, scored and interpreted. Normed technical excellence encompassed evidence regarding test reliability (test-retest, internal consistency and alternate forms), replicability, range of coverage, and score gradation.

The CSE test evaluation team was composed of eleven people--all but one of whom held an advanced degree in some field of education. In each of the four rating categories, they applied a detailed point rating system. Within each category a test or subtest could receive from 0 to 15 points.* Then four letter grades were assigned to each test or subtest.

> Within each of the four evaluative categories, a letter grade based on the points assigned to each aspect of each criterion was computed. The sum of the points, ranging from 0 to 15, was the determinant of the letter grades in the following manner:
>
> "G" (Good; 12 to 15 points). The test meets the criterion very well. Little improvement appears necessary for its immediate utilization. CSE would endorse such a measure or employ it in its own assessment efforts.
>
> "F" (Fair; 8 to 11 points). The instrument is probably among the better tests available, but it does not completely meet the criterion. Alterations would probably render the test a candidate for "G" status. If no better qualified measure were available, CSE would employ an "F" measure, but would interpret results more cautiously.
>
> "P" (Poor; 0 to 7 points). The instrument does not meet the criterion; it is clearly unsatisfactory in that area. Rather than employ such a measure for assessment, it would be better seek an alternate device or attempt to develop one.
>
> Each test then earned four letter grades by the MEAN system. The four-letter combination serves as the Center's official evaluation of the test. (p. xxxi)

*For a detailed account of how the CSE evaluators applied the MEAN evaluation system, see pp. vi-xxxiii of CSE ESTE, 1976. The CSE volume does not give information on the inter-rater reliability of ratings; it does state, though, that:

> The primary concerns in the application of the MEAN system were the objectivity of and consistency of the evaluations. To maximize both the objectivity for any one test evaluator and the consistency across evaluators specific guidelines for evaluation of each criterion and letter-grade assignment were adopted. (p. xix)

Test Evaluation Results

What sort of grades did the tests get from the UCLA evaluators? Table 1 summarizes the results of the CSE test evaluations by educational domain or goal area and by grade level. The first column under each grade level shows simply the total number of tests and subtests categorized as appropriate to that grade level. The second column under each level shows the number of tests and subtests that are rated at least "F", or fair, in each of the four categories of rating. In other words, shown in the second column are the number of tests and subtests rated at least fair or "FFFF" with respect to measurement validity, examinee appropriateness, administrative usability, and normed technical excellence.*

The MEAN evaluation results are nasty. At first and second grade levels, in nine out of fourteen goal areas, not a single test attained an "FFFF" rating. At the 3-4 and 5-6 grade levels, results were slightly better. At these levels somewhat more tests met the "FFFF" criterion. Nevertheless, the proportion of tests rated "FFFF" or better is remarkably low. Across all grade levels and all fourteen educational goal areas, over 90 per cent of all standardized tests were rated as poor or clearly unsatisfactory in at least one area. Not a single test or subtest out of nearly 4000 rated received an unambiguous "GGGG" rating. These findings are sobering. Are all standardized tests this unsatisfactory?

Where Do The Tests Show Up Best?

Table 1 indicates that tests do not receive the same sorts of ratings in all goal areas and at all grade levels. Ratings seem slightly higher at grade levels three through six than at grades one and two. At the first grade level, tests categorized by CSE into the cognitive and intellectual skills area show up as relatively strong. At the second grade level, tests of language arts and mathematics show up somewhat stronger than tests categorized into other areas. At 3-4 and 5-6 grade levels, the types of tests showing up relatively better according to CSE ratings are cognitive and intellectual skills, language arts, mathematics and reading. These are also the areas in which most tests are available (with the notable exception of affective and personality traits tests). Also, most of these domains or goal areas are the traditional territory of educational achievement testing. This pattern suggests that in future work, we ought to take a closer look at the sorts of tests available in these areas. For example, it would be useful to know why tests in these areas seem to receive higher ratings; for technical, historical or other reasons?

*The CSE evaluators do not indicate what relative weights they would give to each of the four evaluative categories in which they rated tests and subtests. However, since they say that a "poor" rating means that an instrument is clearly unsatisfactory, tests meeting the "FFFF" criteria would seem to be at least minimally satisfactory in terms of the CSE ratings. In other words, this manner of summarizing the CSE ratings presumes that a good rating in one evaluative category does not compensate for a poor rating in another category.

TABLE 1: UCLA Center for the Study of Evaluation, Elementary School Test Evaluations

DOMAIN or GOAL AREA	GRADE 1		GRADE 2		GRADES 3-4		GRADES 5-6	
	No. of tests	No. Rated at Least FFFF	No. of tests	No. Rated at Least FFFF	No. of tests	No. Rated at Least FFFF	No. of tests	No. Rated at Least FFFF
1. Affective and Personality Traits	443	0	434	0	633	0	556	0
2. Arts and Crafts	24	0	25	0	28	0	28	0
3. Career Education	25	0	26	0	68	0	74	0
4. Cognitive and Intellectual Skills	366	6	304	3	438	9	408	23
5. Foreign Language	13	0	29	0	33	0	99	0
6. Language Arts	58	0	98	7	199	22	196	22
7. Mathematics	100	1	134	6	286	20	289	20
8. Music	0	0	1	0	61	1	73	4
9. Perceptual and Motor Skills	376	0	286	0	309	0	260	0
10. Physical Education and Health	18	0	23	0	33	0	41	0
11. Reading	468	1	481	5	592	27	426	29
12. Religion and Ethics	11	0	10	0	12	0	11	0
13. Science	8	1	9	0	19	1	22	1
14. Social Studies	18	0	21	0	46	0	51	1

Even in the goal areas where tests show up strongest according to CSE ratings, however, the proportion of tests reaching the "FFFF" level seems remarkably low. In no goal areas at any of the four grade levels does the proportion of tests rated at least "FFFF" reach even fifteen per cent. Why do the tests receive so many "poor" or "clearly unsatisfactory" ratings?

Why Unsatisfactory?

There are several alternative explanations for the findings from the CSE test evaluations. Standardized tests themselves may generally be as unsatisfactory as the CSE findings indicate; but the standards used by the CSE evaluators may have been inappropriate or those standards may have been wrongly applied.*

The latter possibility was raised in a recent study by the Government Accounting Office (GAO) entitled Problems and Needed Improvements in Evaluating Office of Education Programs.** The GAO commented on the CSE test evaluations as follows:

> There was general agreement among the educators and test publishers interviewed that the Center's work was a definite improvement over other existing evaluations of standardized tests. Some test publishers, however, criticized the Center's work because it was based on incomplete data, excluding, for example, the highly technical specifications from which test questions are developed. They also criticized it for establishing educational goals on which test ratings were based without consulting publishers, and for relying on the work of graduate students.
>
> In response, Center officials indicated that they had sent letters to test publishers advising them of the study and its purpose, stating what information was desired, and asking for any other material the test publishers wished to send. They acknowledged that their curricular goal system is not perfect; however, they believe it is a tremendous step forward, providing a clear statement of expected student behaviors which test users can employ to match tests to curriculums. Officials also stated that the system is justified because test makers often measure skills that are different from what they claim to measure. Center officials also noted that all participating graduate students demonstrate adequate competence in research and measurement before being

*For example, the CSE test evaluation system gave no attention to the issue of test costs and little, if any, attention to the issue of cultural bias.

**Washington, D.C.: GAO, 1977.

> hired; they followed set procedures, were routinely monitored, and discussed questionable points with supervisors. State education agency officials and measurement experts interviewed stated that the Center's work should be continually updated and should include an assessment of available criterion-referenced tests (pp. 47-8).

It is not clear, then, that the standards used by CSE in testing the tests, and the manner in which CSE applied those standards, are the sole explanation for their startling findings. Attention must be given to the quality of standardized tests themselves. In response to the question set out at the start of this paper, "How good are standardized tests for the functions they serve?", an initial answer based on the CSE test evaluations seems very often to be "poor" or "clearly unsatisfactory."

The tests do not, however, seem to be equally unsatisfactory in all four evaluative categories. The two categories in which tests most often seemed to be rated unsatisfactory by CSE were measurement validity and normed technical excellence. This pattern suggests that additional attention needs to be given to standards for judging validity of standardized tests and for judging tests' norming and other technical qualities.

Unsatisfactory for What?

The issue of whether tests are satisfactory or unsatisfactory is inevitably bound up with the question of what functions tests serve. The CSE evaluators do not make entirely clear the perspective from which they were testing the tests. Nevertheless, it seems that their test evaluations were carried out mainly from the perspective of using standardized tests for program evaluation. As the foreword to the CSE Elementary School Test Evaluations notes, CSE is "concerned exclusively with issues in the evaluation of social programs." (p. v.). So it was apparently from this perspective that CSE found what they called "a rather bleak picture of the present quality of evaluative instruments" (p. xxxiii).

In fairness, it should be noted that many test publishers likely did not develop their tests specifically as program evaluation instruments. Therefore, they might properly claim that their tests were evaluated by CSE in terms of a function which they were not intended to serve. Nevertheless, the findings of the CSE test evaluators point up a real contradiction; whatever the purposes they were designed for, published standardized tests are widely being used for program evaluation; yet the CSE test evaluations suggest that most published tests are not satisfactory for this function. This contradiction leads to other questions.

Is there any relationship between quality of tests and frequency of their use?

The inquiry by Heyneman and Mintz, cited above, used the CSE criteria to examine whether better tests were used more often. They found that there did appear to be a relationship between the quality of tests and the frequency of their use in federally sponsored research:

> There does appear to be a positive relationship between the quality of tests and their frequency of use. But the degree is not equally as strong from one test category to another. A preference for the better rated instruments is particularly evident with tests of academic achievement. More equivocal results appear in tests of vocational skills and intelligence, though in certain respects researchers are definitely using the better of those available in these two categories. The anomaly lies in the categories of reading tests and tests of personality, where the higher rated of the tests have no better chance of being utilized than those judged to be of poor quality.*

These findings raise several issues. Why does there appear to be a relationship between quality ratings and test use in some areas but not in others? Is there a relationship between rated quality and test use for functions other than federal research? If tests are not selected for use on the basis of quality ratings, how are they chosen? These questions deserve more attention.

What are the implications of lower test quality ratings at the lower grade levels?

Table 1 shows that the CSE test ratings are substantially lower for grade one and grade two tests than for higher grade level tests. Indeed, the only domain in which more than one test received an "FFFF" rating at the first grade level was the cognitive and intellectual skills area. This raises the question of whether program evlauation can be carried out at lower grade levels on bases comparable to evaluations at higher grade levels. More generally this pattern also raises questions regarding the appropriateness of standardized testing with young children. The CSE results are not promising here.

*Heyneman and Mintz, The Frequency and Quality of Measures Utilized in Federally Sponsored Research on Children and Adolescents. The criterion of quality used in the Heyneman and Mintz study was derived from CSE test evaluation ratings in the measurement validity and normed technical excellence categories. Heyneman and Mintz curiously report that "The other two categories [examinee appropriateness and administrative usability] were omitted because they were designed to help the classroom teacher or educational administrator." p. 5.

Will selection of tests on the basis of rated quality tend to bias the sorts of criteria by which programs are evaluated?

Among the fourteen educational goal areas in which tests were rated by CSE, relatively few had more than a handful of tests which received an "FFFF" rating or better. In some categories, for example social studies and arts and crafts, not a single test was rated "FFFF" or better at any of the grade levels. If evaluation instruments are selected on the basis of ratings such as those by CSE, this would seem implicitly to bias criteria for program evaluation away from some education goals, toward reading and math, where the better tests are found. Is this appropriate? If not, how can such bias be avoided?

Are the standards by which CSE rated tests appropriate, and what different standards might be appropriate for judging tests for different functions?

As suggested above, questions have already been raised about the manner in which CSE rated test instruments. Other questions doubtless could also be raised. For example, does it even make sense to rate the quality of test instruments for program evaluation, apart from the nature of programs to be evaluated? Were the criteria by which CSE rated test validity appropriate? How do the standards by which CSE rated tests compare with the APA, AERA, and NCME Standards for Educational and Psychological Tests?* Moreover, how might standards for rating tests' quality vary depending on the functions which tests are to serve?

* * * * *

These questions are important, but have no easy answers. In future staff work for the National Consortium on Testing, we will explore these and other issues regarding the quality and functions of tests used in American education.

Walt Haney and David K. Cohen
Huron Institute
123 Mt. Auburn St.
Cambridge, MA 02138

*Joint Committee of the American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, Standards for Educational and Psychological Tests (Washington, D.C.: APA, 1974).

* * * * * * * * * *

SOURCES OF INFORMATION ON STANDARDIZED TESTS:

There are a variety of sources of information on standardized tests and reviews of tests available. This appendix identifies some of the more prominent sources.

The seven Mental Measurements Yearbooks, all edited by Oscar Krisen Buros, are:

- The Nineteen Thirty-Eight Mental Measurements Yearbook, 1938 (reissued 1972). (This volume is sometimes referred to as The First Mental Measurements Yearbook.)
- The Nineteen Forty Mental Measurements Yearbook, 1941 (reissued 1972). (This volume is sometimes referred to as The Second Mental Measurements Yearbook.)
- The Third Mental Measurements Yearbook, 1949.
- The Fourth Mental Measurements Yearbook, 1953.
- The Fifth Mental Measurements Yearbook, 1959.
- The Sixth Mental Measurements Yearbook, 1965.
- The Seventh Mental Measurements Yearbook, 1972.

In addition, Buros has also published two listings of tests in print which provide indexes to the tests reviewed in the seven MMYs and several other monographs on tests of particular types. These are:

- Test in Print, 1961 (sometimes called Tests in Print I).
- Tests in Print II, 1974.
- English Tests and Reviews, 1975
- Foreign Language Tests and Reviews, 1975.
- Intelligence Tests and Reviews, 1975.
- Mathematics Tests and Reviews, 1975.
- Personality Tests and Reviews II, 1975.
- Reading Tests and Reviews II, 1975.
- Science Tests and Reviews, 1975.

- Social Studies Tests and Reviews, 1975.
- Vocational Tests and Reviews, 1975

All of Buros' books are published by the Gryphon Press of Highland Park, New Jersey.

The CSE test evaluation series published by the UCLA Graduate School of Education includes:

- CSE Elementary Test Evaluations, 1970.
- CSE-ECRC Preschool/Kindergarten Test Evaluations, 1971.
- CSE-RBS Test Evaluations; Tests of Higher-Order Cognitive, Affective, and Interpersonal Skills, 1972.
- CSE Secondary School Test Evaluations, 1974 (3 volumes).
- CSE Elementary School Test Evaluations, 1976.

The most comprehensive source of information on unpublished tests are:

- O.G. Johnson and J. W. Bommarito. Tests and Measurements in Child Development, 1971; and
- O. G. Johnson. Tests and Measurements in Child Development Handbook II, both published by Jessey-Bass of San Francisco.

These are simply some of the more comprehensive sources of information on tests. In future staff circulars, we will provide information on other source material, and in particular on publications by Consortium members concerning testing and alternatives to testing.